PROYECTO - ENTREGA 2

DAVID MEJÍA CASTAÑO, 1007221842
ANDERSON VALENCIA BERMÚDEZ, 1000869230
DIEGO HERNANDO ARANGO RÍOS, 1042768156

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS E INGENIERÍAS / MODELOS Y SIMULACIÓN DE SISTEMAS I

UNIVERSIDAD DE ANTIOQUIA

DOCENTE
RAÚL RAMOS POLLÁN

Predict students' dropout and academic success

Usando el Data set de Kaggle Predict students' dropout and academic success (kaggle.com). Este dataset cuenta con 4425 muestras.

Este conjunto de datos se puede utilizar para comprender y predecir la deserción escolar y los resultados académicos. Los datos incluyen una variedad de factores demográficos, socioeconómicos y de rendimiento académico relacionados con los estudiantes matriculados en instituciones de educación superior. El conjunto de datos proporciona información valiosa sobre los factores que afectan el éxito de los estudiantes y podría usarse para guiar las intervenciones y políticas relacionadas con la retención de estudiantes.
Con este conjunto de datos, los investigadores pueden investigar dos preguntas clave:

- ¿Qué factores predictivos específicos están relacionados con la deserción o finalización de los estudiantes?
- ¿Cómo interactúan las diferentes funciones entre sí?

A continuación, se va a presentar paso a paso para la selección de los datos por medio de este data set.

1) Progreso alcanzado

Eliminar de manera aleatoria un 5% por ciento de datos en 3 distintas columnas (`'Marital status', 'Previous qualification', 'Gender'`). Esto lo logramos con un notebook creado en google colab, de la siguiente manera:

```
[2]  # Importar librerias
     import pandas as pd
     import numpy as np
```

```
[3]  #Cargar Archivo csv
     df = pd.read_csv('/content/dataset.csv')

```

```
     # eliminar datos al azar de las columnas indicadas
     columnas = ['Marital status', 'Previous qualification', 'Gender']
     porcentaje_faltante = 0.05

     for columna in columnas:
         n_filas = len(df)
         n_eliminar = int(n_filas * porcentaje_faltante)
         indices_eliminar = np.random.choice(n_filas, n_eliminar, replace=False)
         df.loc[indices_eliminar, columna] = np.nan

     # guardar el archivo modificado
     df.to_csv('newdataset.csv', index=False)
```

```
[5]  import os
     os.getcwd()
```

```
'/content'
```

2**. Modificar un nuevo Archivo 'newdataset.cvs' ´:** Luego reproducimos con la librería pandas del notebook colab para revisar los valores faltantes en las columnas.

```python
import pandas as pd

# cargar el archivo modificado
df_new = pd.read_csv('/content/newdataset.csv')

# revisar cantidad de valores faltantes en las columnas de interés
columnas = ['Marital status', 'Previous qualification', 'Gender']
for col in columnas:
    n_missing = df_new[col].isna().sum()
    print(f"La columna {col} tiene {n_missing} valores faltantes.")
```

```
La columna Marital status tiene 221 valores faltantes.
La columna Previous qualification tiene 221 valores faltantes.
La columna Gender tiene 429 valores faltantes.
```

3. **Resumen de limpieza de datos**

usamos la función isnull() para identificar los valores faltantes y la función sum() para contarlos. Luego, usamos la función mean() para obtener el porcentaje de valores faltantes por columna y eliminamos aquellas que superen el umbral del 50% con la función dropna(). Finalmente, usamos la función fillna() para completar los valores faltantes en las columnas restantes con la media. Es importante tener en cuenta que hay muchas técnicas diferentes para la imputación de datos y la eliminación de filas y columnas con valores faltantes, y que la elección de la técnica adecuada depende del tipo de datos y del análisis que se desee realizar.

NOTA: cuando se ejecuta df.fillna(df.mean(), inplace=True), los valores faltantes en las columnas del dataframe se completan con la media de cada columna respectiva. Los valores completados se quedan en su posición original en el dataframe, reemplazando los valores faltantes. La opción inplace=True indica que el dataframe original se modificará directamente, en lugar de crear una copia con los valores completados.

```python
#empezar a trabajar en la limpieza de los datos, identificando y completando los valores faltantes
import pandas as pd
import numpy as np

# cargar el archivo CSV
df = pd.read_csv('/content/newdataset.csv')
# mostrar las primeras filas del dataset
print(df.head())
# obtener información general del dataset
print(df.info())
# obtener estadísticas descriptivas del dataset
print(df.describe())
# identificar valores faltantes
print(df.isnull().sum())
# obtener porcentaje de valores faltantes por columna
print(df.isnull().mean() * 100)
# eliminar columnas con más del 50% de valores faltantes
umbral = 0.5
df.dropna(thresh=umbral*len(df), axis=1, inplace=True)
# completar los valores faltantes en las columnas restantes con la media
df.fillna(df.mean(), inplace=True)
```

3. Construir el modelo

Para el avance de la segunda entrega hemos explorados unos datos categóricos y numéricos para correlación de variables teniendo en cuenta que tan dependientes o independientes hay entre ellas. Luego seleccionar la que mejor precisión se acerca para la construcción de nuestro modelo. Por lo tanto, hemos llegado hasta ese punto y mas adelante vamos a seguir explorando datos de importancia para la mejor significancia que dé como mejor resultado para el modelo y de acuerdo con esos resultados le daremos un análisis inferencial del tema de la deserción estudiantil que hay en diferentes planteles educativos el cual entrenaremos datos del 80% que continúan y 20% de prueba para estudiantes que desean desertar.

Cabe aclarar que nuestro objetivo principal es desarrollar un modelo de Machine Learning capaz de predecir si un estudiante dado continuará o abandonará sus estudios en la educación superior. Esto podría ayudar a las instituciones educativas a identificar a los estudiantes en riesgo y tomar medidas proactivas para retenerlos.

Y que uno de nuestro criterio deseable es analizar datos importantes (demográficos y socioeconómicos) de los estudiantes y que se correlacionen con su rendimiento académico de esta manera se visualizan patrones y tendencias el cual logre predecir si continúa o no continua en el programa o área de interés que cursa actualmente.

La métrica con la cual será analizado el rendimiento del modelo es Área bajo la curva ROC (AUC): la medida de la capacidad del modelo para distinguir entre los estudiantes que abandonan o tienen éxito académico y los que no lo hacen.

REFERENCIAS

https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention.